OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1–10

doi: 10.1093/scan/nsw143
Original Article

Brain activation for spontaneous and explicit false belief tasks overlaps: new fMRI evidence on belief processing and violation of expectation

Lara Bardi,[1] Charlotte Desmet,[1] Annabel Nijhof,[2] Jan R. Wiersema[2], and Marcel Brass[1]

[1]Department of Experimental Psychology, and [2]Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

Correspondence should be addressed to Lara Bardi, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: lara.bardi@ugent.be.

Abstract

There is extensive discussion on whether spontaneous and explicit forms of ToM are based on the same cognitive/neural mechanisms or rather reflect qualitatively different processes. For the first time, we analyzed the BOLD signal for false belief processing by directly comparing spontaneous and explicit ToM task versions. In both versions, participants watched videos of a scene including an agent who acquires a true or false belief about the location of an object (belief formation phase). At the end of the movies (outcome phase), participants had to react to the presence of the object. During the belief formation phase, greater activity was found for false *vs* true belief trials in the right posterior parietal cortex. The ROI analysis of the right temporo-parietal junction (TPJ), confirmed this observation. Moreover, the anterior medial prefrontal cortex (aMPFC) was active during the outcome phase, being sensitive to violation of both the participant's and agent's expectations about the location of the object. Activity in the TPJ and aMPFC was not modulated by the spontaneous/explicit task. Overall, these data show that neural mechanisms for spontaneous and explicit ToM overlap. Interestingly, a dissociation between TPJ and aMPFC for belief tracking and outcome evaluation, respectively, was also found.

Key words: fMRI; spontaneous theory of mind; false belief

Introduction

Our capacity to represent others' mental states, goals, beliefs and intentions is of paramount importance for successful social interaction. In traditional theory of mind (ToM; Premack and Woodruff, 1978) tasks participants are required to explicitly reason about the other's mental states. In the 'Sally–Anne' false-belief task, Sally observes an object being placed in a box and then leaves the room. Following this, Anne moves the object to a different box. When Sally reenters the room, participants must indicate the location, in which they think Sally will look for the object, that is thus they must represent Sally's false belief (Wimmer and Perner, 1983). Based on this research, ToM has been characterized as a hallmark of human cognitive development: the attribution of mental states to others requires executive resources (i.e. the ability to inhibit one's own perspective) and language resources and emerges relatively late in the development. Only by the age of 4 years, children are able to pass false belief tasks (Wimmer and Perner, 1983; Baron-Cohen *et al.*, 1985; McKinnon and Moscovitch, 2007; Wellman *et al.*, 2001; Gweon *et al.*, 2012).

Recently a mounting body of evidence has challenged this traditional view. Behavioral tasks, measuring reaction times or spontaneous looking patterns in adults and infants, suggest that the ability to track beliefs and even false beliefs of others may be engaged spontaneously in adults (Senju *et al.*, 2009; Kovács *et al.*, 2010; Schneider *et al.*, 2011) and present already in infants, before children are able to pass standard false belief

Received: 19 April 2016; **Revised:** 30 August 2016; **Accepted:** 21 September 2016

tasks (Clements and Perner, 1994; Onishi and Baillargeon, 2005; Southgate *et al.*, 2007; Surian *et al.*, 2007; Kovács *et al.*, 2010; Senju *et al.*, 2011). This research shows that we represent other's beliefs even when we are not required to do so and even in situations where other's mental states are completely irrelevant for our current goals.

In the study by Kovács *et al.* (2010), 7-month-old infants are presented with a video representing an agent who obtains certain knowledge about the location of an object. After that, infants display surprise, as indexed by increased looking time, when they are presented with a picture of the object that is contrary to the agent's belief (Kovács *et al.*, 2010). In this study, the same paradigm was also applied to adults, who showed a similar effect; in their case, it was indicated by their reaction times to the (expected or unexpected) presence of the object. After the presentation of a scene, reaction times to the appearance of an object were short not only when the participant expected the object to be present but also when the agent only (false belief condition) believed the object would be present. Critically, participants were never asked to consider the agent's belief. Schneider *et al.* (2011) provided evidence for such spontaneous belief processing using a false-belief anticipatory looking paradigm. Here, participants observed some movie clips depicting an agent having true or false beliefs about the location of an object. At the end of the movies, participants had eye movement patterns consistent with belief tracking even though they reported not to have been consciously engaged in mentalizing. This suggests that people spontaneously engage in belief tracking as if they were explicitly asked to do it. Stronger support for the overlap between spontaneous and explicit mentalizing comes from a recent study. In their elegantly designed work, Schneider *et al.* (2014a) investigated the extent to which the operation of spontaneous ToM is modulated by task instructions. One group of participants was given no task instructions, another was instructed to track the position of the ball in the scene, and a third was asked to do a tracking of the agent's belief. Despite different task goals, all groups' eye-movement patterns were consistent with belief analysis.

These findings represent evidence that humans spontaneously track the belief states of others in an unintentional manner.

However, the question remains whether spontaneous and explicit forms of ToM are based on the same cognitive/neural mechanisms or rather reflect qualitatively different processes. While some authors have generally questioned that spontaneous ToM tasks reflects mentalizing of any kind (Heyes 2014; Phillips *et al.*, 2015), others (Apperly and Butterfill, 2009) have proposed two distinct ToM systems. They suggest the spontaneous ToM system is present early in life, is fast and efficient and operates spontaneously/unconsciously whereas the explicit form would develop later and would be slower, more deliberate and flexible, but therefore also more cognitively demanding. Finally, Carruthers (2016) postulates just a single mindreading system, which sometimes operates fully automatically, sometimes in conjunction with the standing goal of anticipating people's behavior, and sometimes in a more controlled way (by involving executive function and working memory).

Despite this lively debate, studies investigating similarities and differences between spontaneous and explicit ToM are still scarce (Schneider *et al.*, 2014b; Van der Wel *et al.*, 2014; Rosenblau *et al.*, 2015). To date, neuroimaging studies have mostly focused on the neural correlates of explicit belief processing (e.g. Fletcher *et al.*, 1995; Gallagher *et al.*, 2000; Ruby and Decety, 2003; Saxe and Kanwisher, 2003). These studies have revealed a quite consistent pattern of brain regions involved when participants are asked to reason about somebody else's false belief. What is referred to as the 'ToM network' includes the temporo-parietal junction (TPJ), medial prefrontal cortex (MPFC), superior temporal sulcus (STS) and precuneus (PC). A number of recent meta-analysis studies confirm that ToM, across different tasks, consistently activates TPJ and MPFC (e.g. Decety and Lamm, 2007; Van Overwalle, 2009; Schurz *et al.*, 2014).

To the best of our knowledge, only one study investigated the neural bases of spontaneous ToM alone (Kovács *et al.*, 2014) and only two studies have compared brain activation for spontaneous and explicit ToM in the same participant with conflicting results (Schneider *et al.*, 2014b; Hyde *et al.*, 2015). In the study of Schneider *et al.* (2014b), brain activity was measured both during the spontaneous ToM task of Schneider *et al.* (2011) described above and a classical explicit task based on the presentation of a text describing short stories. The authors first identified a set of ROIs through an explicit localizer task where participants had to read short stories and answer questions about a person's belief. Following this, they observed that only a subset of the regions showed significant activation for false beliefs in the spontaneous task. In particular, the left STS and posterior cingulate (PC), showed the predicted pattern (false belief > true belief) during the spontaneous ToM video clips, while TPJ did not. This outcome contrasts with the results of Hyde *et al.* (2015) who, using Near-Infrared-Spectroscopy, found significant activation in the TPJ ROI during spontaneous false belief task.

However, it is important to note that both studies used spontaneous and explicit tasks that involved different stimulus materials and different procedures, making it difficult to interpret possible similarities/differences in brain activation.

The aim of the current study was therefore to compare brain activity related to spontaneous and explicit mentalizing directly, using a within-subjects design and identical stimuli and procedure. To this end, participants, during fMRI, were presented with a new developed task. In this task, participants watch short movies depicting an object moving in the scene and an agent forming a true or false belief about the location of the object in the outcome (belief formation phase; Kovács *et al.*, 2010; Deschrijver *et al.*, 2015). Participants are instructed to respond to the presence of the object in the outcome phase at the end of the movie (object detection). The new procedure integrates catch questions presented at the end of the movies in a small percentage of the trials so that the agent's belief either remained irrelevant for the task (spontaneous version; the questions concerned the color of the agent's cap) or was relevant (explicit version; participants were interrogated about the agent's belief). After the spontaneous task, that was always presented first, a debriefing session was included to ensure that participants were unaware of the belief manipulation.

We compared brain activity during the belief-tracking phase for false and true belief conditions in the spontaneous and explicit ToM tasks. Furthermore, we also looked at violation of expectation in the outcome phase. If the other's beliefs are spontaneously represented on-line, the evaluation of an outcome will be affected by both the belief of the participant and the belief of the agent.

Method

Participants

Twenty-three healthy students (5 males; age: mean = 22, ranging from 19 to 25) participated on the basis of written informed

consent. One participant was excluded from the final analyses due to an error in data saving. The study was conducted according to the Declaration of Helsinki, with approval of the local ethics committee of the University Hospital Gent. All subjects had normal or corrected-to-normal vision. No subject had a history of neurological, major medical or psychiatric disorder. All participants were right-handed as assessed by the Edinburgh handedness questionnaire (Oldfield, 1971).

Procedure and design

The experiment comprised two main parts presented in a fixed order: an spontaneous ToM task and an explicit version of the same task. The two versions of the ToM task were identical except for the presentation of catch questions at the end of some trials. The catch questions were included to distract participants from the belief manipulation and to induce active belief processing in the spontaneous and explicit task, respectively (see below). The entire testing session was limited to 1 hour. Participants were lying in the MRI scanner while watching short videos via a mirror. Our stimuli were created based on the study of Kovács *et al.* (2010).

All movies consisted of two phases: the belief formation phase and the outcome phase. The movies in the belief formation phase differed along two aspects of the belief attributable to the agent (Buzz Lyghtyear from the cartoon Toy Story): the agent's belief could be true or false (true: matching reality and participant knowledge; false: not matching reality and participant's knowledge) and belief content (positive content: the agent believes the ball is present; negative content: the agent believes the ball is absent). The presence or the absence of the ball in the outcome phase was completely independent of the belief formation phase, because the ball was randomly present in 50% of the trials in all the conditions (see below). Combined with the two versions of the outcome phase (ball does or does not appear from behind the occluder), there were 8 different conditions (8 movies) and movies were repeated 10 times in a random order for each task version resulting in a total of 160 experimental trials. Responses were given through a response box.

Participants kept their right and left index and middle fingers on different buttons. The experiment consisted of two sessions in which the spontaneous and the explicit ToM versions were presented. Each version lasted about 25 min and consisted of 2 separate blocks (fMRI runs) with a short break in between. After completion of the spontaneous version of the ToM task, participants filled in a debriefing form based on the one used by Schneider *et al.* (2013), which was adapted to the current task and translated to Dutch. It consisted of five questions (see Appendix A for an English translation). By the use of this form we checked whether participants were aware of our belief manipulation.

Stimuli and task

All movies comprised a belief formation phase and an output phase. Each movie lasted 13.8 s.

Belief formation phase. As shown in Figure 1, all movies started with an agent placing a ball on a table in front of an occluder. Then the ball rolled behind the occluder. Following this, the movies could continue in four ways depending on the experimental conditions: (i) In the True Belief-Positive Content condition (P+ A+), the ball rolled out of the scene from behind the occluder, and then rolled back behind the occluder (ball last seen by the participant at 10 s; time information is given relative to the beginning of the movie) in the agent's presence. The agent left the scene at 11 s. Thus, the agent could rightly believe the ball to be behind the occluder. (ii) In the True Belief-Negative Content condition (P− A−), the ball emerged from behind the occluder without leaving the scene, then rolled back behind the occluder, and finally left the scene (ball last seen at 10 s), all in the agent's presence. The agent left the scene at 11 s. Thus, the agent could rightly believe the ball not to be behind the occluder. (iii) In the False Belief-Positive Content condition (P− A+), the order of when the ball and the agent left the scene was reversed relative to the True Belief-Negative Content condition. Thus, the agent left the scene at 6 s. Then, the ball emerged from behind the occluder without leaving the scene, rolled back behind the occluder, and finally left the scene (ball last seen at 11 s), all in the agent's absence. Thus, the agent could wrongly believe the ball to be behind the occluder. (iv) In the False Belief-Negative Content condition (P+ A−), the ball rolled out of the scene from behind the occluder in the agent's presence. Then, the agent left the scene at 9 s. In his absence, the ball rolled back behind the occluder at 11 s. Thus, the agent could wrongly believe the ball not to be behind the occluder. As in the original task, in order to keep participants' attention during the presentation of the movies, they were instructed to press a key with the index finger of their left hand when the agent left the scene.

Outcome phase. At the end of each movie, the agent re-entered the scene and the occluder fell down. The four conditions were paired with two equally probable outcomes, in which the ball was either present or absent behind the occluder. Participants were instructed to press a key as fast as possible with the index finger of their right hand when they detected the ball. The presence or the absence of the ball was completely independent of the belief formation phase, because the ball was randomly present in 50% of the trials in all the conditions. As a result of the combination of belief formation phase (P− A−, P+ A+, P+ A−, P− A+) and output phase (ball present B+, ball absent B−) there were eight different movies. Each movie was repeated 10 times. Therefore, the entire experiment comprised 80 trials for the spontaneous version and 80 for the explicit version.

Spontaneous/explicit manipulation. The movies were identical for the spontaneous and explicit versions. During both the spontaneous and the explicit task, a question appeared randomly in 18 trials after the end of the movie. Questions were presented in black text on a light grey background for 1000 ms. In the spontaneous version, the question was: 'Did Buzz have a blue cap?' The cap could be either blue (50% of the movies) or red (50%). In the explicit version, the question was: 'Did Buzz think the ball was behind the screen?' This was also true in 50% of the movies. Questions were presented with a variable jitter interval. A pseudo-logarithmic jitter was applied. Half of the inter-trial intervals were short (range between 200 and 2000 ms steps of 600 ms), one-third was intermediate (range between 2600 and 4400 ms) and one-sixth was long (range between 5000 and 6800 ms) with a mean inter-trial interval of 2700 ms.

The words 'Yes' and 'No' were presented on the left or right of the screen in both task versions. A 50% of catch questions had 'Yes' printed left and 'No' right, 50% *vice versa*. In this way, responses could not be planned in advance. Participants had to respond to the answer on the left with their left middle finger and to the answer on the right with their left index finger.

fMRI data acquisition

Images were collected with a 3 T Magnetom Trio MRI scanner system (Siemens Medical Systems, Erlangen, Germany) using a

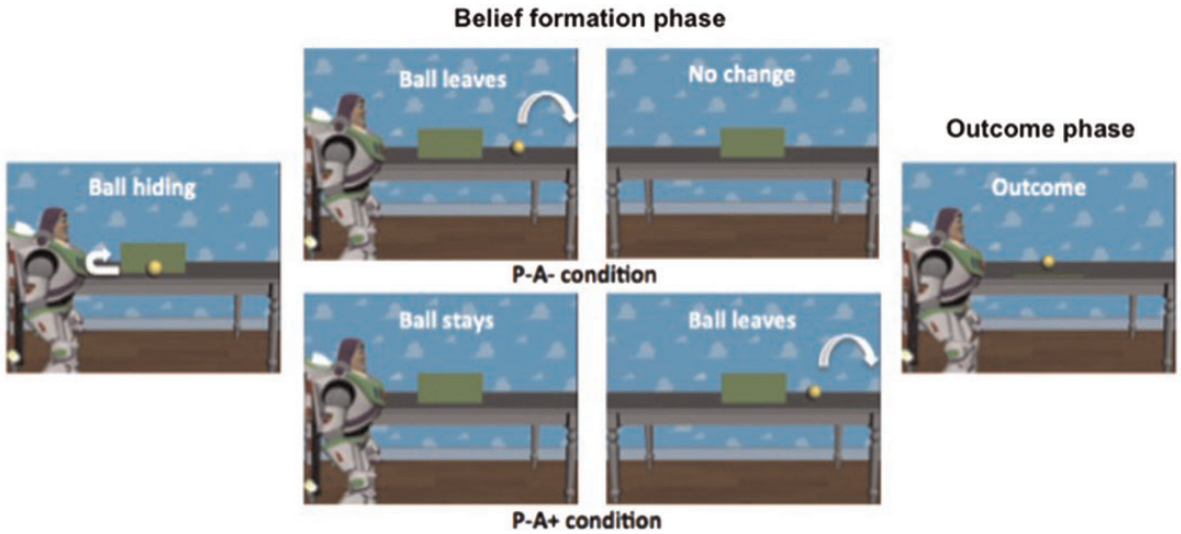

Fig. 1. Frames of two of the movies presented during the tasks. Example of a false belief condition (P–A−) and a true belief condition (P–A+). There were eight conditions in total, resulting from the combination of belief formation phase and outcome phase. In the first part of the movie, the ball rolls behind the screen. In the second part (belief formation phase), in the presence of the agent, the ball can change location or stay behind the occluder. Afterwards the agent leaves the scene and the ball can change its location or not. In the outcome phase, the agent comes back to the scene and the occluder is lowered. The ball is present or not (50% of the cases). Please note that, in the 'no change' video fragment, the ball was moving anyway. For example, it would roll out from the occluder and then roll back behind the occluder. In all movies, the ball was visible to the participant for the same amount of time.

32-channel radiofrequency head coil. Before the experiment started, 176 high-resolution anatomical images were acquired using a T1-weighted 3D MPRAGE sequence [repetition time (TR) = 2530 ms, echo time (TE) = 2.58 ms, image matrix = 256 × 256, field of view (FOV) = 220 mm, flip angle = 78°, slice thickness = 0.90 mm, voxel size = 0.9 × 0.86 × 0.86 mm (resized to 1 × 1 × 1 mm)]. Next, the experiment was performed during which whole brain functional images were obtained. The functional images were acquired using a T2*-weighted EPI sequence sensitive to BOLD contrast (TR = 2000 ms, TE = 28 ms, image matrix = 64 × 64, FOV= 224 mm, flip angle = 80°, slice thickness = 3.0 mm, distance factor = 17%, voxel size = 3.5 × 3.5 × 3.0, 34 axial slices). Volumes were aligned along the AC–PC axis.

fMRI data preprocessing

The fMRI data were analyzed with SPM8 software (Wellcome Department of Cognitive Neurology, London, UK). The first four volumes of all EPI series were used to allow the magnetization to approach a dynamic equilibrium and were excluded from the analysis. Data preprocessing started with spatially realigning the functional images using a rigid body transformation. Then the realigned images were slice time corrected with respect to the first slice. The high-resolution structural image of each subject was co-registered with the mean image of the EPI series. During segmentation, the structural scans were brought in line with the tissue probability maps available in SPM. The parameters estimated during the segmentation step were then used to normalize the functional images to standard MNI space. Finally, the functional images were resampled into 3 × 3 mm voxels and spatially smoothed with a Gaussian kernel of 8 mm (full-width at half maximum).

Behavioral data analysis

In both versions, spontaneous and explicit, we recorded RTs for the detection of the ball at the end of each movie. Behavioral performance therefore reflects a spontaneous measure of ToM in both cases. Behavioral data were analyzed with IBM SPSS Statistics 20 (SPSS, Inc., Chicago, IL, USA). For one participant, detection responses were not correctly recorded due to technical problems (therefore, his/her data were excluded from this analysis). However, the participant performed correctly on the catch questions, showing that he/she was engaged in the task. We performed a repeated-measures ANOVA on reaction times, with task (spontaneous, explicit), belief (true belief, false belief) and belief content (positive content, negative content) as within-subject factors.

fMRI data analysis

The subject-level statistical analyses were performed using the general linear model. The model contained separate regressors for all possible combinations of Belief (true belief, false belief) and Belief Content (positive content, negative content) for the belief formation phase (duration of 9 s from the moment in which the agent places the ball on the table to the moment in which the agent comes back to the scene). For the outcome phase, there were separate regressors for all possible combinations of Belief, Belief Content and Outcome (ball present, ball absent) (duration of 0 s). In total, the model included 12 regressors of interest for the spontaneous task and 12 regressors of interest for the explicit task. Six subject-specific regressors that were obtained during the realignment step were added to account for head motion. All resulting vectors were convolved with the canonical hemodynamic response function to form the main regressors in the design matrix (the regression model). The statistical parameter estimates were computed separately for each voxel for all columns in the design matrix. Contrast images of interest were created at the first level and were then entered into a second level analysis with subject as a random variable. Contrasts at this group level were made using one-sample *t*-tests.

Contrasts were run separately for the belief formation phase and the outcome phase. For the belief formation phase, in order to identify regions involved in false belief tracking (belief formation

phase), our main contrast of interest was computed as follow: false belief P− A+ and P+ A− (participant's and agent's belief do not match) > true belief P− A− and P+A+ (participant's and agent's beliefs match). The interaction between belief and task (explicit > spontaneous task) was also calculated as follow: (P− A+ and P+ A− explicit > P− A− and P+ A+ explicit) > (P− A− and P+ A+ spontaneous > P− A+ and P+ A− spontaneous). In addition, we calculated a contrast based on the content of the agent's belief as follows: A+ > A−.

For the outcome phase, the agent's and the participant's belief about the presence of the ball were considered with respect to the actual presence of the ball in the outcome phase (B− ball absent, B+ ball present). Therefore, we analyzed separately conditions B− and B+ conditions. Violation of expectation was calculated as a mismatch of the belief content and the actual presence of the ball in the output phase. For example, if the ball was absent at the end of the movie (B−), the agent's expectation (only) would be violated in the P− A+ condition. To identify regions involved in the violation of expectation, we computed the main contrasts of interest as follows: (i) violation of expectation based on the agent's belief was calculated as follow: P− A+ > P− A− for B− trials and P+ A− > P+ A+ for B+ trials. (ii) Violation of expectation based on participant's belief: P+ A− > P− A− for outcomes with no ball (B−) and P− A+ > P+ A+ for outcomes with ball (B+). We also computed the interaction between agent's violation of expectation and task: spontaneous > explicit and explicit > spontaneous and between participant's violation of expectation and task (spontaneous > explicit and explicit > spontaneous). To correct for multiple comparisons a cluster-extent based thresholding approach was used (Friston *et al.*, 1996). First a primary uncorrected threshold of $P < 0.001$ at voxel level was used to identify groups of suprathreshold voxels. Second, a cluster-level extent threshold, represented in units of contiguous voxels (k), was determined by SPM 8 ($P < 0.05$ FWE cluster corrected threshold). Only clusters that have a k value that is equal or larger than this threshold are reported. The coordinates reported correspond to the MNI coordinate system.

In addition to the main analyses, we carried out a signal-change analysis in the *a priori* defined region of interest (ROI) based on a meta-analysis of peaks reported in 26 studies on mentalizing (see Kovács *et al.*, 2014). The ROI was a sphere with a radius of 10 mm centered on the coordinates 56 −47 33. Mean β's for the events of interest were extracted using the MARSBAR toolbox for SPM (Brett *et al.*, 2002). The β values obtained were then subjected to a repeated-measures ANOVA containing the factors task (spontaneous task, explicit task), belief (false belief, true belief) and belief content (positive content, negative content).

Results

Behavioral results

Performance on the catch question was well above chance both for the spontaneous and the explicit version, with an accuracy level of 82% (range of correct responses: 7–18) and 74% (range of correct responses: 4–18), respectively. For the ball detection task, a significant main effect of belief was found [$F(1, 20) = 14.9$, $P < 0.05$, $\eta^2 = 0.43$), with RTs in false belief conditions being faster than RTs in true belief conditions. More importantly, we found a significant interaction between belief and belief content [$F(1, 20) = 7.94$, $P < 0.05$, $\eta^2 = 0.28$). Pairwise comparisons of conditions revealed that participants were significantly slower in P− A− trials than in all other trials (Ps < 0.05). This pattern of results overlaps completely with data from the original paper of Kovács *et al.* (2010). The fact that participants are faster both when they expected the ball to be behind the occluder and when only the agent expected the ball to be behind the occlude (P− A+ condition) confirms that participants spontaneously represent the other's belief during the detection task. Interestingly, although spontaneous and explicit versions exhibited a very similar pattern of RTs, the impact of the agent's belief on performance, was even slightly stronger in the spontaneous version as attested by the interaction between task and belief content [$F(1, 20) = 4.4$, $P = 0.049$, $\eta^2 = 0.18$). In the spontaneous task version, the difference in RTs between positive content trials (A+) and negative content trials (A−) was larger than in the explicit version. This reveals that the effect of the agent belief's content was even stronger in the spontaneous than in the explicit version. No interaction between belief, content and task was found. Overall, this outcome supports the idea that, in spontaneous version, beliefs are spontaneously processed. Therefore, task instructions requiring participants to explicitly attend the other's perspective do not induce qualitative changes in behavioral performance (Figure 2). This pattern of data completely overlaps with the results obtained in a recent behavioral study from our group using the same stimulus material (Nijhof *et al.*, submitted for publication) and with what has been previously shown with a different task (Schneider *et al.*, 2014a).

fMRI results

First, we aimed to identify regions that were involved in false belief processing (false belief). Concerning the belief formation phase, higher activity during false belief than true belief occurred in angular gyrus (AG) (peak coordinates: 42 −67 43) and in fusiform gyrus/collateral sulcus (peak coordinates: 33 −52 1). Importantly, the computation of the interaction between belief and task did not lead to any significant cluster of activation. Also, no clusters emerged for the contrast positive > negative belief content. Concerning the outcome phase, for outcomes where the ball was absent (B−), violation of expectation based on participant's belief (P+ A− > P− A−, B−) revealed activation in the anterior MPFC (peak coordinates: 9 38 7). Violation of expectation based on the agent's belief (P− A+ > P− A−, B−) also activates the aMPFC (peak coordinates: 6 38 −8). The AG and the aMPFC clusters are shown in Figure 3. Moreover, the left anterior insula (peak of activation: −33 32 −11), the right occipital cortex, the left fusiform gyrus and the right sensorimotor cortex were activated (Table 1 for a complete list of the areas and coordinates). The anterior insula has been associated with the evaluation of the affective consequences of our action (Brass and Haggard, 2010; Koban and Pourtois, 2014) and it has been considered crucial for the social interaction (e.g. Cracco *et al.*, 2016). For outcomes where the ball was present (B+), violation of expectation based on participant's belief (P− A+ > P+ A+, B+) revealed activation in the supplementary motor area (SMA), right sensorimotor cortex and the occipital cortex. Violation of expectation based on the agent's belief (P+ A− > P+ A+, B+) did not lead to any cluster of activation. Importantly, neither agent's violation of expectation nor for the participant did the interaction between expectation and task reveal any result. Motor activation emerging in different contrasts probably reflects motor preparation of the left-side response that was needed to respond to the catch questions. Although the catch questions were only presented in a small percentage of trials (about 20%), participants could not predict in which

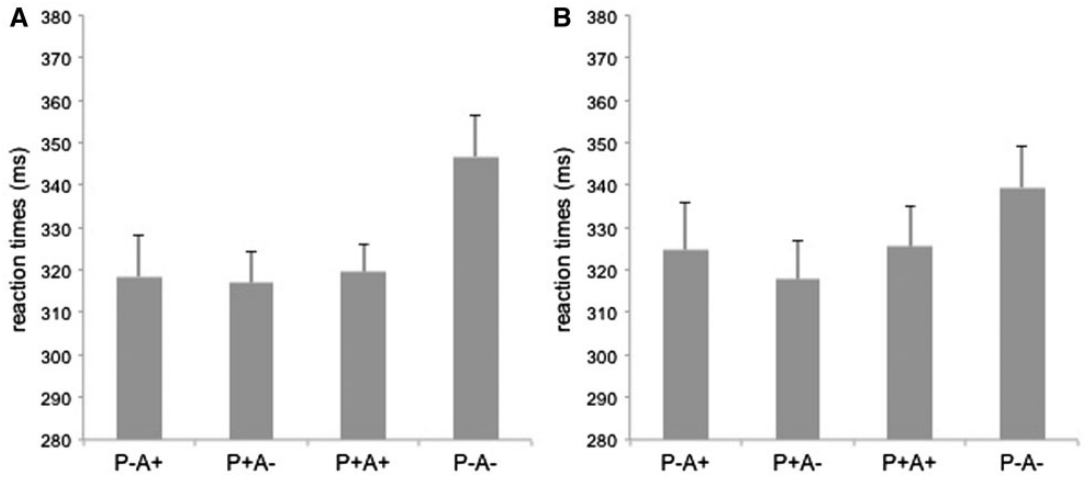

Fig. 2. Reaction times for ball detection (spontaneous measure) are displayed for the four conditions of the task. (A) Behavioral performance under spontaneous ToM task instructions. (B) Behavioral performance under explicit task instructions.

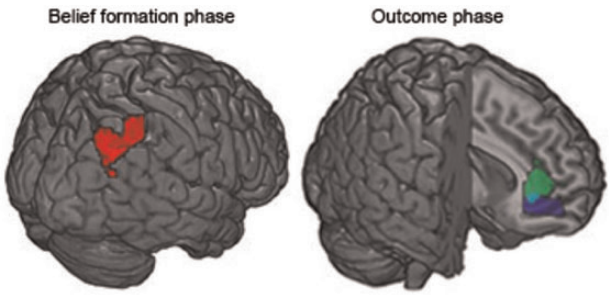

Fig. 3. *Left panel.* Cluster of activation in the PPC for the contrast false belief > true belief (irrespective of the task) during the belief formation phase. *Right panel.* Clusters of activation in the MPFC for the participant's violation of expectation (participant positive content prediction > negative outcome in green) and the agent's violation of expectation (agent positive content prediction > negative outcome in blue) (irrespective of the task) in the outcome phase.

trials they would be present and therefore always prepared the response. Results are summarized in Table 1.

ROI analysis. With this analysis we wanted to further explore potential differences in belief processing between spontaneous and explicit ToM versions of the task. More specifically, in the study of Kovács *et al.* (2014), an asymmetric effect has been found in activation of the right TPJ during the spontaneous ToM version. Results showed that TPJ is only active when a false belief attributed to the agent has a positive content (the agent thinks that the ball is behind the occluder). Such an asymmetry may be due to task instructions, which required participants to respond only to the presence of the target, but not to its absence. An alternative explanation is that this asymmetry is a functional characteristic of the spontaneous belief tracking system, which leads to preferential encoding of certain types of belief contents, while ignoring others in specific situations. If the asymmetry noticed by Kovacs *et al.* (2014) only occurs in the spontaneous task then this would suggest that only spontaneous processing has this specialization.

We found a main effect of task [$F(1, 21) = 4.39$, $P < 0.05$, $\eta^2 = 0.17$) with higher activation in the spontaneous when compared with the explicit version. This effect is attributable to a general decrease of the BOLD signal across the task blocks and should not be interpreted as a specific effect of task manipulation on TPJ activity.

Table 1. Peaks of activation from different contrasts in the belief formation phase and outcome phase of the videos

Area	MNI peak coordinates xyz	Cluster size	Z-scores
Belief formation phase False > true belief			
Angular gyrus	42 −67 43	197	5.13
Fusiform gyrus/collateral sulcus	33 −52 1	149	7.15
Outcome phase. Ball absent (B−)			
A + P − > A− P −			
aMPFC	6 38 −8	112	4.28
Right sensorimotor	42 −22 46	536	5.34
Right occipital cortex	45 −82 −8	443	4.67
Left occipital cortex	−15 −64 −2	152	3.78
Left anterior insula	−33 32 −11	228	4.28
A + P − > A− P −			
aMPFC	9 38 7	292	4.24
Outcome phase. Ball present (B+)			
A + P − > A + P +			
SMA	−3 −7 58	180	4.00
Thalamus	−6 −13 −5	649	5.25
Right sensorimotor	39 −16 43	508	5.09
Right occipital	36 −94 −8	653	4.52
Lingual gyrus	−12 −67 −8	177	4.49
Left occipital	−42 −79 −8	286	4.44

More important, there was a main effect of belief with higher activation values for false than for true belief conditions [$F(1, 21) = 4.81$, $P < 0.05$, $\eta^2 = 0.17$). Importantly, no interaction between belief and task emerged, supporting the results on our whole brain analysis. Furthermore, there was a main effect of content [$F(1, 21) = 4.62$, $P < 0.05$, $\eta^2 = 0.18$] with higher activation when the agent belief has a positive content (Ball+). Importantly, a significant interaction effect of belief and content was found [$F(1, 21) = 7.98$, $P < 0.05$, $\eta^2 = 0.27$]. Post hoc comparisons revealed a significantly higher activation for the false belief, positive content condition as compared to the false belief, negative content ($P < 0.05$). However, no interaction with the task emerges. Our data therefore support the idea that our task is sensitive to the content of false belief. However there is no evidence for a dissociation between the spontaneous and explicit version. Percentages signal change has been depicted in Figure 4 for all task conditions.

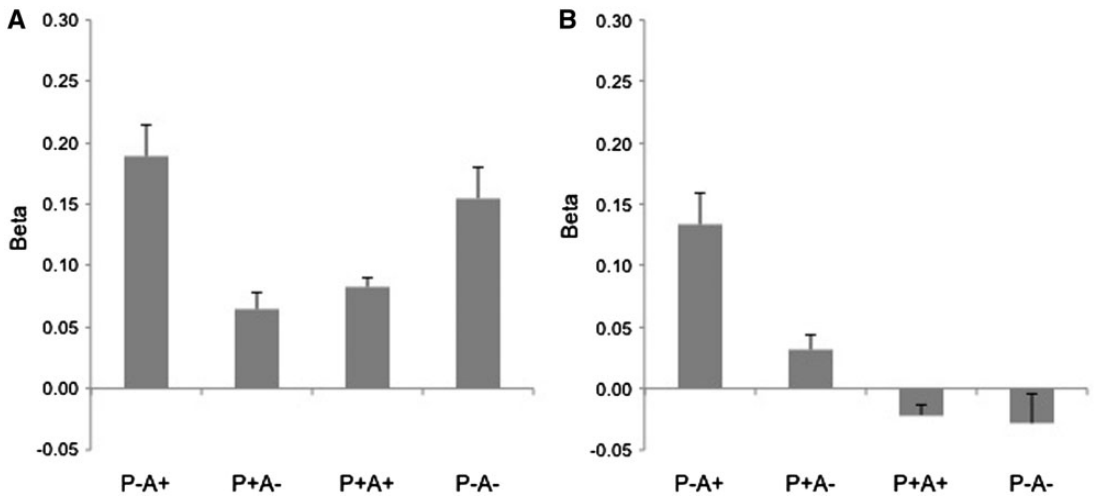

Fig. 4. Percentages signal change (Beta signal change) in the TPJ ROI for all task conditions are depicted.

Discussion

To the best of our knowledge, this is the first study that compares spontaneous and explicit false beliefs processing by adopting the same task procedure and stimuli in both task versions. Our data strongly support the idea that during spontaneous and explicit ToM, we track others' belief by using the same neural mechanisms that have been shown to be involved in explicit mentalizing. Importantly, the two task versions only differed for the spontaneous/explicit processing of the agent's belief. A debriefing procedure, allowed us to verify that participants were unaware of the belief manipulation during the spontaneous version of the ToM task. On the other hand, in the explicit version, we have ensured that participants would explicitly track the agent's belief. When we analyzed behavioral results from the ball detection task, we have shown the pattern of reaction times is not affected by task instructions (spontaneous *vs* explicit). This outcome support previous observations derived from the same task (Nijhof *et al.*, in press) or a different one using eye gaze as dependent variable (Schneider *et al.*, 2014a). In the study of Schneider *et al.* (2014a), the authors gave participants either no instruction, the instruction to track the position of the ball in the scene, or the instruction to track the agent's belief. Here, eye-movement patterns were consistent with belief analysis, irrespective of the task.

A possible limitation of our design is the fixed order of the tasks. The spontaneous ToM task had to come always first as participants would otherwise become aware of the belief manipulation. Although a general decrease of participants' attention to the stimuli is expected in the second part of the experiment, it is important to note that responses to catch questions in the explicit task were still well above chance level (74%), showing that participants were engaged in the task.

Our results show that spontaneous ToM processing activates the same neural network that has been previously pointed out as being critically involved in explicit ToM, specifically the right TPJ and anterior MPFC (e.g. Fletcher *et al.*, 1995; Gallagher *et al.*, 2000; Ruby and Decety, 2003; Saxe and Kanwisher, 2003). A number of recent meta-analysis studies indicate that explicit ToM, across different tasks, consistently activates TPJ and MPFC (e.g. Decety and Lamm, 2007; Van Overwalle, 2009; Schurz *et al.*, 2014). During the belief formation phase, greater activity for false *vs* true belief conditions was found in the right TPJ. Importantly, no interaction was found with the spontaneous/explicit task. Overlapping results were obtained for the whole-brain analysis (AG activation) and the ROI analysis when an *a priori* TPJ ROI was defined based on a meta-analysis on explicit ToM. Moreover, the MPFC was active during the outcome phase. The MPFC showed greater activation when the outcome did not match expectations based on the preceding events contained in the movie, as compared to when the outcome did match expectations. Interesting, two partially overlapping clusters were found for participant and agent's violation of expectation. Again, no interaction was found with the task version.

Our outcome on the TPJ is in line with previous data from ROIs analyses on a spontaneous ToM task (Kovács *et al.*, 2014; Hyde *et al.*, 2015). Moreover, activation of the TPJ in false belief trials seems to be higher when the belief of the agent has a positive content (i.e. the agent falsely believed the ball was behind the occluder). This pattern is in line with what was found in a previous study adopting the spontaneous version of the same task (Kovács *et al.*, 2010). In the same vein, violation of expectation based on the agent's belief, led to stronger activation in the MPFC in trials with ball absent outcomes (that is, when the agent expected the ball to be behind the occluder, positive content belief). This bias for agent's beliefs with positive content can reflect the specific instructions of our detection task. In fact, participants were asked to respond only when the ball was present at the end of the movie (ball present outcomes). This can explain why the belief of the agent is more salient when it has a positive content, which is when the agent expected the ball to be behind the occluder. On the other hand, this effect might reflect a content-dependent representational constraint, or limit, on spontaneous ToM, which restricts the system to tracking false beliefs that may favor potential behaviorally relevant beliefs. This may reflect a functional difference between spontaneous and *explicit* ToM (Kovács *et al.*, 2014). A possible representational limit of the spontaneous ToM system has been previously identified for object identity (Low and Watts, 2013). However, since responses in our ball detection task reflect a spontaneous measure of ToM in both our task, future studies are needed to give a more definitive answer to this question.

It has been proposed that behavioral effects in spontaneous ToM tasks do not reflect spontaneous ToM but could rather be explained by domain-general cognitive mechanisms, such as response selection, attentional orienting or spatial coding of

response locations, which simulate the effects of mentalizing (Heyes, 2014; Philips *et al.*, 2015). On the contrary, if a spontaneous ToM task reflects mentalizing, one would expect task manipulation to induce activation in brain areas that are commonly associated with ToM (i.e. TPJ and aMPFC). Our fMRI data, together with previous observations (Kovács *et al.*, 2014; Hyde *et al.*, 2015) support this idea. Moreover, Deschrijver *et al.* (2015) recently carried out the same paradigm (spontaneous version only) in a group of adults with autism spectrum disorder (A). Here a 'ToM index' has been calculated as the difference between the P− A− and the P− A+ conditions, representing the degree to which the agent's belief about the presence of the ball influences RTs. The size of individuals' 'ToM index' was found to correlate with A symptom severity in the A group.

Our results are in contrast with the previous neuroimaging study of Schneider *et al.* (2014b). There, having identified ROIs based on an explicit ToM task, during the spontaneous ToM task, the authors found a significant difference in the BOLD signal between false and true belief conditions only in the left STS and precuneus (PC) but not in the TPJ, although the comparison was in the right direction (higher activation for false *vs* true belief). One possible explanation for this discrepancy lies in differences in the tasks and stimuli used. In effect, the explicit task adopted in the study of Schneider *et al.* was a common task used in false belief research involving reading stories describing someone's knowledge and beliefs (linguistic material) and then answering a question. Moreover, in that study, the analysis of the BOLD signal has not been performed on the belief formation phase (referred to as belief set-up sequence) but only on the output phase (belief test phase). In our experiment, TPJ activation only emerges in the belief formation, and not in the outcome phase. Therefore, methodological and analysis differences between the two studies can account for the discrepancy. Finally, there is another, perhaps theoretically more interesting, reason for why our results (and the results of Hyde *et al.*, 2015) differ from those obtained by Schneider *et al.* (2014b). This difference concerns the true belief condition used to compute the main contrast of interest (FB > TB). In the true belief condition of our task, since both the agent and the participant observe the ball reaching its last position before the agent leaves the scene (no changes occur after that), the agent has all the information the participant has. In other words, in the participant's perspective, the agent has knowledge about the position of the ball. In the true belief condition of Schneider *et al.* (2014b), the agent does not have all relevant knowledge, but instead holds a belief that accidentally becomes true at the end. In effect, the ball still moves from its location when the agent is not in the scene and reaches its final position only at the end of the movie. Although it is a matter of debate whether true beliefs are processed differently from the state of reality, or knowledge (Sommer *et al.*, 2007; Aichhorn *et al.*, 2009; Back and Apperly, 2010), both behavioral and neuroimaging evidence suggests that true belief reasoning is different from reasoning about the state of reality. For example, in a study by Döhnel *et al.* (2012), brain activity for false and true belief reasoning has been compared with state of reality-control conditions. When compared with this control condition, right TPJ activity was observed both for true and false belief reasoning. We can argue that the probability of representing true beliefs is higher if true beliefs do not match the state of reality; and therefore, do not completely overlap with the participant's knowledge. In this sense, contrasting a false with a true belief condition (equal to knowledge) as in the present study (and in the study of Hyde *et al.*, 2015) would be more sensitive in capturing belief processing than contrasting a false belief with a true belief (different from knowledge) condition as in Schneider *et al.* (2014b). In any case, further studies are needed to understand whether and how true belief tracking occurs in spontaneous ToM tasks.

Although we are aware that the time resolution of fMRI does not allow us to argue about a strong dissociation between the belief formation phase and the outcome phase in our task, separate regressors for the two phases have been included in the model. Our results suggest that TPJ is more involved in the belief formation phase while the aMPFC seems to be implicated to a greater extent in the outcome phase. Despite extensive research on the role of TPJ and aMPFC in ToM and social cognition in general, the differential role of these two areas is still poorly understood. However, from a functional–anatomical point of view, it is unlikely that both areas serve the same function. For example, Saxe and Powell (2006) suggest that the MPFC recruitment is not restricted to reasoning about another person's thoughts (the later-developing component of ToM) or even subjective, internal states in general, but may be involved more broadly in representing socially or emotionally relevant information about another person. Van Overwalle (2009) hypothesized that the mPFC is engaged in making inferences about permanent social and psychological properties of others, such as personality traits. Moreover, outside the domain of social cognition, the aMPFC has long been associated with the encoding of the outcome (O'Doherty *et al.*, 2002; Kennerley and Wallis, 2009; Rushworth *et al.*, 2011).

In the current work, we have shown neural activation patterns for spontaneous and explicit ToM overlap. However, this does not provide definitive evidence that spontaneous ToM is based on the same mechanisms involved in explicit ToM. We cannot rule out the possibility that activation in the spontaneous condition was driven by attentional processes in the TPJ, causing an overlap with activity for explicit ToM. This interpretational challenge is a limitation for the present and numerous fMRI-based studies in the social neuroscience domain, as interpretations rely on reverse inference (Poldrack, 2006).

The TPJ has been related to key computations in the social domain, such as ToM, self-other distinction in the control of imitation, agency processing and perspective taking (e.g. Ruby and Decety, 2001; Blanke *et al.*, 2002; Farrer and Frith, 2002; Farrer *et al.*, 2003; Saxe and Wexler, 2005; Legrand and Ruby, 2009; Brass *et al.*, 2009) but also in other non-social processes, such as spatial attention (Corbetta *et al.*, 2002; Mitchell, 2008). Although it is out of the scope of this work to enter the discussion about domain-specific or domain-general neural computation of the TPJ, recent models try to reconcile observations from the social and other cognitive neuroscience fields suggesting that TPJ is involved in reorienting of attention toward unexpected relevant events (Corbetta *et al.*, 2008) or 'contextual updating, updating of internal models based on incoming incongruent information' (Geng and Vossel, 2013). Accordingly, TPJ would help updating mental representations based on changes (we did not necessary attend to) occurring in the environment. In line with this idea, we have found preferential activation of TPJ during the tracking period of other's beliefs (belief formation phase). Moreover, following the same idea, we should not expect differences between spontaneous and explicit ToM. Moreover, this outcome is in line with the self-other distinction model were TPJ detects an incongruence between internally generated representation and externally triggered representation (e.g. Brass *et al.*, 2009).

Recently, the temporal profile of processing another person's visual perspective has been investigated using event-related

potentials (ERPs; McCleery *et al.*, 2011). In this study, participants were asked to actively take the perspective of an agent or simply their own regarding the number of displayed disks on a wall. McCleery *et al.* concluded that early in visual perspective processing the temporal and parietal cortices distinguish between self and other perspectives and then, later, the frontal cortex resolves conflicts between these representations during response selection. Importantly, however, McCleery *et al.*'s study design was focused on visual perspective taking and did not examine brain regions involved in a classic false-belief processing task. Moreover, source analysis of the EEG signal referred to the later prefrontal cortex. Although behavioral studies have suggested a possible dissociation between belief calculation/tracking and response selection in ToM (Leslie and Thaiss, 1992; Leslie *et al.*, 2005; Qureshi *et al.*, 2010), it remains unclear what the specific role of TPJ and MPFC is and how these regions may interact with each other.

In summary, our findings suggest that mechanisms underlying the spontaneous tracking of others' beliefs may exploit similar representational systems as explicit ToM judgments do. In fact, neural mechanisms for spontaneous and explicit ToM overlap when tasks are equal in terms of stimulus materials and procedure. Interestingly, the analyses on the belief formation phase and outcome phase suggest dissociation between TPJ and aMPFC, with TPJ being preferentially activated for belief tracking and the aMPFC for outcome evaluation.

Acknowledgment

We would like to deeply thank the reviewers for their fruitful comments on the manuscript.

Funding

This study has been funded by Research Foundation – Flanders (FWO) Pegasus Fellowship to Lara Bardi and grant 331323-Mirroring and ToM, Marie Curie Fellowship (Marie Curie Intra-European fellowship for career development) to L.B.

Conflict of interest. None declared.

References

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. *Journal of Cognitive Neuroscience*, **21**(6), 1179–92.

Apperly, I.A., Butterfill, S.A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, **116**(4), 953–70.

Back, E., Apperly, I.A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, **115**(1), 54–70.

Baron-Cohen, S., Leslie, A.M., Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, **21**(1), 37–46.

Blanke, O., Ortigue, S., Landis, T., Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, **419**, 269–70.

Brass, M., Ruby, P., Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **364**(1528), 2359–67.

Brass, M., Haggard, P. (2010). The hidden side of intentional action: the role of the anterior insular cortex. *Brain Structure and Function*, **214**, 603–10.

Brett, M., Anton, J., Valabregue, R., Poline, J. (2002). Region of interest analysis using an SPM toolbox, Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 26, 2002, Sendai, Japan, Available on CD-ROM in Neuroimage **16** (2).

Carruthers, P. (2016). Two systems for mindreading? *Philosophy and Psychology*, **7**, 141–62.

Clements, W.A., Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, **9**, 377–95.

Corbetta, M., Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, **3**, 201–15.

Corbetta, M., Patel, G.H., Shulman, G.L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, **58**, 306–24.

Cracco, E., Desmet, C., Brass, M. (2016). When your error becomes my error: anterior insula activation in response to observed errors is modulated by agency. *Social Cognitive & Affective Neuroscience*, **11**, 357–66.

Decety, J., Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The Neuroscientist*, **13**(6), 580–93.

Deschrijver, E., Bardi, L., Wiersema, J.R., Brass, M. (2015). Spontaneous theory of mind in adults with autism spectrum disorder: autistic traits predict lesser belief attribution to others. *Cognitive Neuroscience*, **7**, 192–202.

Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., Sommer, M. (2012). Functional activity of the right temporoparietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage*, **60**(3), 1652–61.

Farrer, C., Franck, N., Georgieff, N., Frith, C.D., Decety, J., Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. *NeuroImage*, **18**, 324–33.

Farrer, C., Frith, C.D. (2002). Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. *NeuroImage*, **15**, 596–603.

Fletcher, P.C., Happé, F., Frith, U., *et al.* (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, **57**, 109–28.

Friston, K.J., Holmes, A., Poline, J.B., Price, C.J., Frith, C.D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, **4**, 223–35.

Gallagher, H.L., Happé, F., Brunswick, N., Fletcher, P.C., Frith, U., Frith, C.D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, **38**, 11–21.

Geng, J.J., Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: contextual updating? *Neuroscience and Biobehavioral Reviews*, **37**(10 Pt 2), 2608–20.

Gweon, H., Dodell-Feder, D., Bedny, M., Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development*, **83**, 1853–68.

Heyes, C. (2014). Submentalizing: I'm not really reading your mind. *Psychological Science*, **9**, 121–43.

Hyde, D.C., Aparicio Betancourt, M., Simon, C.E. (2015). Human temporal-parietal junction spontaneously tracks others' beliefs: a functional near-infrared spectroscopy study. *Human Brain Mapping*, **36**, 4831–46.

Kennerley, S.W., Wallis, J.D. (2009). Encoding of reward and space during a working memory task in the orbitofrontal cortex and anterior cingulate sulcus. *Journal of Neurophysiology*, **102**(6), 3352–64.

Legrand, D., Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroima- ging results. *Psychological Review*, **116**, 252–82.

Koban, L., Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neuroscience and Biobehavioral Reviews*, **46**, 1–14.

Kovács, ÁM., Téglás, E., Endress, A.D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science (New York, N.Y.)*, **330**(6012), 1830–4.

Kovács, ÁM., Kühn, S., Gergely, G., Csibra, G., Brass, M. (2014). Are all beliefs equal? Spontaneous belief attributions recruiting core brain regions of theory of mind. *PLoS ONE*, **9**(9), e106558.

Leslie, A.M., Thaiss, L. (1992). Domain specificity in conceptual development: neuropsychological evidence from autism. *Cognition*, **43**(3), 225–51.

Leslie, A.M., German, T.P., Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, **50**(1), 45–85.

Low, J., Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, **24**(3), 305–11.

McCleery, J.P., Surtees, D.R., AGraham, K.A., Richards, J.E., Apperly, I.A. (2011). The neural and cognitive time course of theory of mind. *Journal of Neuroscience*, **31**(36), 12849–54.

McKinnon, M.C., Moscovitch, M. (2007). Domain-general contributions to social reasoning: theory of mind and deontic reasoning re-explored. *Cognition*, **102**, 179–218.

Nijhof, A.D., Brass, M., Bardi, L., Wiersema, J.R. (in press). Measuring mentalizing ability: a within-subject comparison between an explicit and spontaneous version of a ball detection task. *PLos ONE*.

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, **18**, 262–71.

O'Doherty, J.P., Deichmann, R., Critchley, H.D., Dolan, R.J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, **33**(5), 815–26.

Onishi, K.H., Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.)*, **308**(5719), 255–8.

Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, **9**, 97–113.

Phillips, J., Ong, D.C., Surtees, A.D.R., *et al.* (2015). A second look at automatic theory of mind: reconsidering Kovacs, Teglas, and Endress (2010). *Psychological Science*, **26**, 1–15.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, **10**(2), 59–63.

Premack, D., Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, **1**(4), 515–26.

Qureshi, A.W., Apperly, I.A., Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: evidence from a dual-task study of adults. *Cognition*, **117**(2), 230–6.

Rosenblau, G., Kliemann, D., Heekeren, H.R., Dziobek, I. (2015). Approximating spontaneous and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. *Journal of Autism and Developmental Disorders*, **45**, 953–65.

Ruby, P., Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience*, **4**, 546–50.

Ruby, P., Decety, J. (2003). What you believe versus what you think they believe: a neuroimaging study of conceptual perspective-taking. *European Journal of Neuroscience*, **17**(11), 2475–80.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., Behrens, T.E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, **70**(6), 1054–69.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage*, **19**, 1835–42.

Saxe, R., Powell, L.J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science*, **17**, 692–9.

Saxe, R., Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, **43**, 1391–9.

Schneider, D., Bayliss, A.P., Becker, S.I., Dux, P.E. (2011). Eye movements reveal sustained spontaneous processing of others' mental states. *Journal of Experimental Psychology: General*, **141**(3), 433–8.

Schneider, D., Slaughter, V.P., Bayliss, A.P., Dux, P.E. (2013). A temporally sustained spontaneous theory of mind deficit in autism spectrum disorders. *Cognition*, **129**(2), 410–7.

Schneider, D., Nott, Z.E., Dux, P.E. (2014a). Task instructions and spontaneous theory of mind. *Cognition*, **133**(1), 43–7.

Schneider, D., Slaughter, V.P., Becker, S.I., Dux, P.E. (2014b). Spontaneous false-belief processing in the human brain. *NeuroImage*, **101**, 268–75.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, **42**, 9–34.

Senju, A., Southgate, V., White, S., Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science (New York, N.Y.)*, **325**(5942), 883–5.

Senju, A., Southgate, V., Snape, C., Leonard, M., Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science: A Journal of the American Psychological Society/APS*, **22**(7), 878–80.

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage*, **35**(3), 1378–84.

Southgate, V., Senju, a., Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, **18**(7), 587–92.

Surian, L., Caldi, S., Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, **18**(7), 580–6.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, **30**(3), 829–58.

van der Wel, R.P.R.D., Sebanz, N., Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, **130**(1), 128–33.

Wellman, H.M., Cross, D., Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, **72**(3), 655–84.

Wimmer, H., Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**, 103–28.

Appendix. Debriefing form (translated from Dutch)

1. Do you have an idea what the goal of this experiment was?
2. Did you notice anything unusual about the movies?
3. Did you notice any particular pattern or theme to the movies?
4. Did you have any particular goal or strategy?